**Olá, estudante!**

Ao analisar conjuntos de dados, geralmente, você se deparar com atributos categóricos, que podem assumir grande quantidade de valores (categorias) possíveis. Por exemplo, em dados de produção de uma indústria, o atributo Fornecedor pode ter muitos fornecedores possíveis; em dados de saúde pública de um estado com dados sobre postos de saúde, o atributo Nome do Posto também pode ter muitas possibilidades. Por conta dessas ocorrências, quando há uma grande quantidade de valores possíveis, os gráficos de barras são muito bons para analisar dados categóricos, mas eles também podem tornar, principalmente, a visualização de valores menores mais difícil. Pensando nesse caso, uma alternativa para esse problema é a nuvem de palavras, que se adequa a grandes quantidades de categorias.

Portanto, nesta aula, você aprenderá a criar e a customizar gráficos de nuvem de palavras, utilizando a plataforma Kibana.

**Objetivos**

* Aprender a criar nuvens de palavras com a plataforma Kibana;
* Aprender a customizar nuvens de palavras para várias situações;
* Conhecer as vantagens e as desvantagens de nuvens de palavras em relação a gráficos de barras.

## Tópico 1 – Criando gráficos de nuvem de palavras com Kibana

**OBJETIVOS**

* Aprender a utilizar a ferramenta Tag Cloud para criar nuvens de palavras com a plataforma Kibana;
* Associar diferentes métricas à nuvem de palavras;
* Aprender a utilizar diferentes escalas de tamanhos de fonte na ferramenta Kibana;
* Comparar as vantagens e as desvantagens entre os gráficos de barras e as nuvens de palavras.

Como você já deve saber, o gráfico de barras é uma ótima abordagem para visualizar atributos categóricos. Porém, à medida que aumenta o número de valores possíveis para as categorias, a visualização do gráfico de barras vai sendo prejudicada. Pode-se constatar, então, que ele não escala bem, ou seja, essa ferramenta não tem um bom resultado, quando é necessário apresentar muitos valores diferentes. Pensando nessa problemática, uma alternativa eficiente e bastante utilizada é a nuvem de palavras.

Esse gráfico, chamado em inglês de tag cloud ou word cloud, é um tipo de visualização que mostra uma coleção de palavras, termos ou frases curtas dispostas em um agrupamento, ao formato de nuvem, em que o tamanho da palavra é proporcional a sua importância. Na plataforma Kibana, a importância de um certo dado está representada por qualquer métrica que você desejar associar a partir dos dados, como média ou soma de valores.

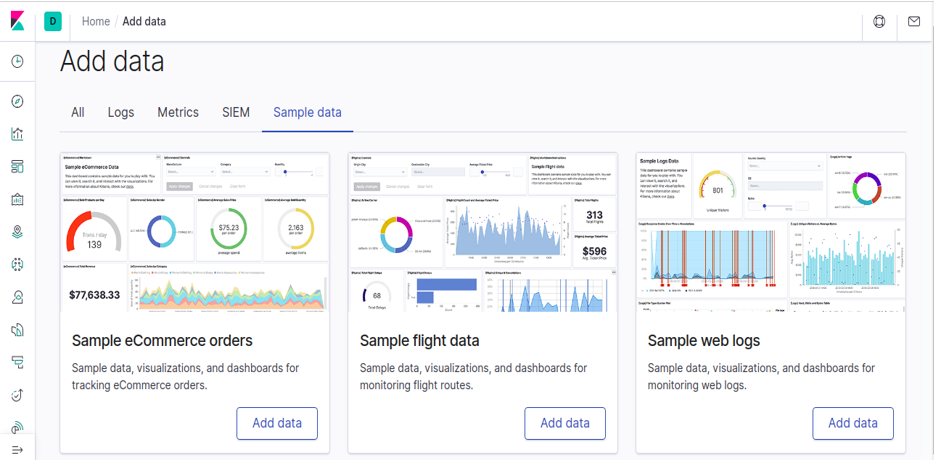
Neste tópico, você deverá aprender a como criar nuvens de palavras, utilizando a ferramenta Kibana, além de saber como associar diferentes métricas às palavras, além de modificar a escala e a orientação das palavras e, por fim, comparar as vantagens e as desvantagens desse tipo de esquema em relação ao gráfico de barras.

Bons estudos!

Você já deve ter alguma familiaridade com o Kibana, a plataforma Web de visualização de dados integrada ao mecanismo de busca Elasticsearch, e deve saber que ele permite a criação de diversos tipos de visualizações que ajudam a explorar os conjuntos de dados dinamicamente. Então, para ilustrar a utilidade das nuvens de palavras, vamos utilizar um exemplo do conjunto de dados sobre voos, oferecido pelo Kibana, como amostra.

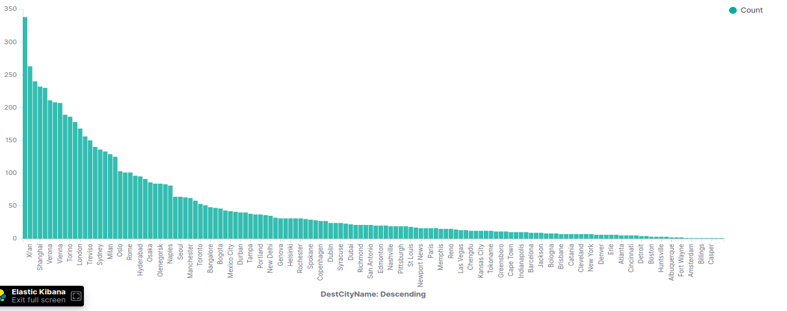
Para acompanhar os exemplos desta aula, inicialize o Elasticsearch e o Kibana. Depois, abra a página dessa plataforma no seu navegador, usando este endereço <http://localhost:5601/>. Agora, caso você não tenha implantado os dados de voo no Kibana, é aconselhável fazê-lo agora.

No centro da página inicial da ferramenta instalada, selecione a opção **Load a data set and a Kibana dashboard** e, no painel **Sample flight data**, acione o botão **Add Data**, como é possível conferir na figura a seguir:

Figura 1 - Carregando dados de exemplo

Esses dados contêm diversas informações sobre voos do mundo todo, incluindo as cidades de destino de cada voo. Como você pôde observar, há uma grande quantidade de cidades que foram destino de voos. Portanto, se quiséssemos visualizar o número de vezes que cada cidade foi destino de um voo, e, assim, conhecer as mais visitadas ou as menos visitadas, uma primeira abordagem poderia ser a utilização de um gráfico de barras, concorda? Pois bem, iremos testar essa opção a seguir.

Em seguida, verifique a figura representada por um gráfico de barras criado com o Kibana, mostrando a quantidade de vezes que cada cidade foi destino de um voo ao longo de um ano:

Figura 2 - gráfico de barras da quantidade de vezes que cada cidade foi destino de voos.

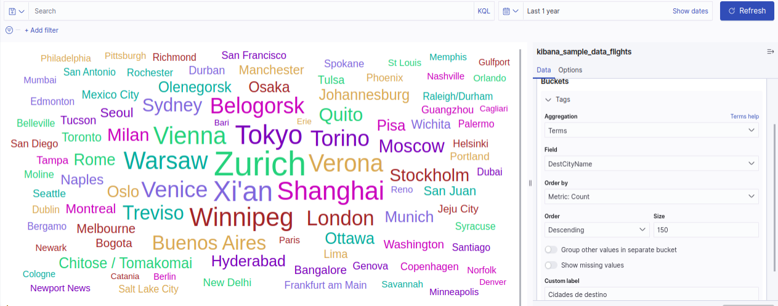
Note que o gráfico tem alguns problemas: primeiro, como a quantidade de cidades é muito grande, foi preciso diminuir a espessura das barras para que o gráfico coubesse na tela, dificultando a identificação do valor associado a cada cidade; segundo, por consequência das inúmeras barras, na legenda do eixo x, o nome de várias cidades foram omitidos por questão de espaço; terceiro, por fim, as barras que representam valores menores, praticamente desaparecem devido à escala relativa entre as cidades com maior número de voos e aquelas com quantidade de voos mais baixa.

A partir desses empecilhos observados, perceba que é muito difícil ter uma noção geral desses dados através do gráfico de barras, por mais interativo que seja o esquema, e por mais que você ainda tenha a possibilidade de utilizar o mouse para obter informações sobre cada barra. Logo após, vamos criar uma nuvem de palavras para tentar melhorar essa visualização, preparado?

Bem, como já foi mencionado, a nuvem de palavras exibe um agrupamento de palavras, em que o tamanho da fonte de cada palavra é proporcional a alguma métrica, geralmente, à quantidade de vezes que a palavra ocorre. Para isso, o Kibana utiliza a conhecida lógica de Metrics e Buckets: em Metrics, são selecionados os valores a serem exibidos; e em Buckets, são selecionadas as formas que esses valores serão agrupados. Para criar uma nuvem de palavras que mostre quais cidades tiveram mais destinos de voo, siga este passo a passo:

1. Abra a página de visualizações, acionando o ícone **Visualize** na barra de navegação;
2. Selecione o botão **Create visualization** e escolha a opção **Tag cloud**;
3. Acione o conjunto de dados de voo: kibana\_sample\_data\_flights;
4. Configure a janela de tempo para **Last 1 year**;
5. Note que, no painel Metric, em **Tag size**, está selecionada a agregação Count, indicando o tamanho de cada palavra associado à contagem, ou seja, a quantas vezes ela ocorre. Inicialmente, você terá acesso apenas à palavra “all” no centro da visualização. Isso ocorre, porque os dados não estão agrupados em torno de um atributo. Para realizar tal ação, no painel **Bucket**, selecione a opção **Add** e selecione **Tags**;
6. Em **Aggregation**, selecione **Terms** e acione o atributo DestCityName, ou seja, as tags (palavras) indicarão as cidades de destino dos voos;
7. Configure **Order** e **Size** para Descending e 150, respectivamente, indicando que serão mostradas as 150 cidades mais visitadas;
8. Configure **Custom label**, como Cidades de destino;
9. Escolha o botão **Update** para atualizar a imagem. O resultado deverá ser similar ao da figura da próxima página;
10. Salve a visualização, selecionando o botão **Save** e escolhendo um nome apropriado.

Tudo bem até aqui? Ótimo! Vamos para a próxima página, para verificar como o resultado deve ficar.

Figura 3 - nuvem de palavras da quantidade de vezes que cada cidade foi destino de voos

Primeiro, analisando que as cidades com mais destinos de voo possuem um tamanho de fonte maior e estão localizadas no centro, enquanto as cidades com menos voos são menores e tendem a ser posicionadas na borda da “nuvem”; perceba também que, no canto inferior esquerdo da tela, há um ícone de alerta. Portanto, colocando o mouse sobre esse ícone, ou selecionando-o com o teclado, aparece uma tooltip informando que, devido ao tamanho da janela, a nuvem de palavras não pode ser exibida completamente, pois ela corta algumas palavras.

Por mais que isso aconteça, é possível consertarmos, quando você adicionar essa visualização a um dashboard e configurar o tamanho do painel, permitindo que ela mostre todas as informações.

Na próxima página, confira uma figura que representa melhor essa ação.

Figura 4 - nuvem de palavras visualizada em um painel maior.

Perceba que, comparada ao gráfico de barras, a nuvem de palavras escala melhor para uma grande quantidade de categorias, dando uma visão geral sobre a ocorrência de cada palavra. Ao contrário da maioria das visualizações, esses tipos de esquema se tornam visualmente melhores quando há mais palavras para mostrar. Além disso, são bastante intuitivas de se compreender. Por esses motivos, observe também que a cor não representa informação alguma nessa nuvem de palavras, possuindo apenas uma função estética.

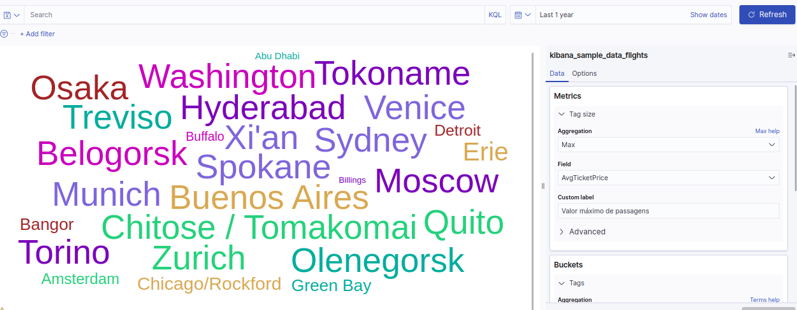
Nesse exemplo, a nuvem mostra a contagem de voos. Portanto, podemos verificar que Zurique, Tóquio, Xi’an e Xangai foram cidades com destino de muitos voos. Enquanto Mumbai, Nova Orleans e Cidade do Cabo foram localidades que tiveram poucos destinos de voo.

Apesar dessa amostragem, é possível fazer uma nuvem em que o tamanho de cada palavra esteja associado a outra métrica, que não a contagem de ocorrências. Ficou curioso? Verifique isso a seguir!

Bem, e se você quiser que o tamanho das tags esteja associado ao preço máximo da passagem dos voos para uma determinada cidade? Simples, basta modificar a agregação utilizada no painel Metrics. Portanto, para exemplificar essa escolha, siga este passo a passo:

1. Na página de visualizações, acione o botão **Create visualization**, selecione a opção **Tag cloud** e o conjunto de dados de voos: kibana\_sample\_data\_flights;
2. No painel **Metrics**, expanda **Tag size** e, em **Aggregation**, escolha Max;
3. Em **Field**, acione AvgTicketPrice. Isso significa que o tamanho das tags estará associado ao valor máximo das passagens dos voos com destino a cada cidade;
4. No painel **Bucket**, selecione a opção **Add** e escolha **Tags**;
5. Em **Aggregation**, acione **Terms** e selecione o atributo DestCityName;
6. Configure **Order** e **Size** para Descending e 150, respectivamente, indicando que serão apresentadas as 150 cidades mais visitadas;
7. Configure **Custom label**, como Cidades de destino.
8. Escolha o botão **Update** para atualizar a imagem. O resultado deverá ser similar ao da disponível na próxima página;
9. Salve a visualização, acionando a opção **Save** e selecionando um nome apropriado.

Então, tudo bem até aqui? A seguir, confira a figura que representa o resultado do passo a passo.

Figura 5 - nuvem de palavras das cidades com voos mais caros.

Nesse gráfico, observe que algumas das cidades que tiveram passagens mais caras foram Spokane, Xi'an e Buenos Aires. Novamente, devido ao tamanho do painel, algumas tags são cortadas, mas isso é corrigido, adicionando a visualização a um dashboard.

Além da métrica utilizada para o tamanho das tags, o Kibana provê outras opções para customizar a nuvem de palavras, que você conhecerá a seguir.

Nesse ponto, estando no painel de edição, acione o botão **Options** no lado superior do painel. Feito isso, surgem novas opções para configurar o estilo da nuvem de palavras, conforme ilustrado na figura a seguir:

Figura 6 - opções de customização de nuvens de palavras

No campo **Text scale**, você pode escolher a escala que irá mapear o valor da métrica para o tamanho da fonte. O padrão é a escala linear (Linear), que mapeia o valor da métrica para o tamanho da fonte da tag de forma uniforme. Por exemplo, se um dado possui o valor da métrica 10 vezes maior que outro, então o tamanho da tag do primeiro será 10 vezes maior que a do segundo na escala linear. Alternativamente, pode-se utilizar a escala logarítmica, reduzindo as diferenças nos valores das métricas.

Na prática, isso significa que a diferença no tamanho da fonte será menor para valores mais altos. Por exemplo, se um dado possui o valor da métrica 10 vezes maior que outro, então, na escala logarítimica, o tamanho da tag do primeiro será apenas 2 vezes maior que a do segundo. Por fim, use a escala da raiz quadrada (Square root), caso você deseje exagerar as diferenças no valor da métrica. Ou seja, o significado dessa ação vai representar a diferença no tamanho da fonte, como sendo menor para valores mais baixos. Portanto, experimente usar essas diferentes escalas.

Contudo, quando não há grande diferença entre os valores da métrica, as diferenças entre as escalas podem não ser evidente.

Além da escala, a plataforma Kibana fornece opções para modificar a inclinação das tags no campo **Orientations**. O padrão é Single, em que todas as tags estão na mesma orientação horizontal. Além disso, também pode-se escolher Right angled, que dispõe algumas tags na horizontal e outras na vertical (em ângulo de 90°); e opção Multiple, que dispõe as tags em diversos ângulos, formando uma visualização “bagunçada”.

Por fim, pode-se também configurar o intervalo de tamanho, em que as fontes das tags podem assumir em **Font size range in pixels**. Portanto, todas essas configurações devem ser escolhidas cuidadosamente, de acordo com o objetivo da visualização. Tudo entendido até aqui? Pois bem, discutiremos isso um pouco mais a seguir.

Apesar das vantagens da nuvem de palavras sobre o gráfico de barras que já foram discutidas, a tag cloud também possui importantes limitações, que devem ser levadas em consideração, quando forem utilizadas.

A exemplo desses empecilhos, tem-se o fato de palavras com mais letras receberem mais ênfase do que vocábulos com menos letras, mesmo que a métrica associada a elas tenha o mesmo valor, pois as palavras com mais letras ocupam, naturalmente, mais espaço. Por exemplo, relembre o modelo da nuvem de palavras com os preços máximos de passagens ilustrado, de novo, na figura a seguir:

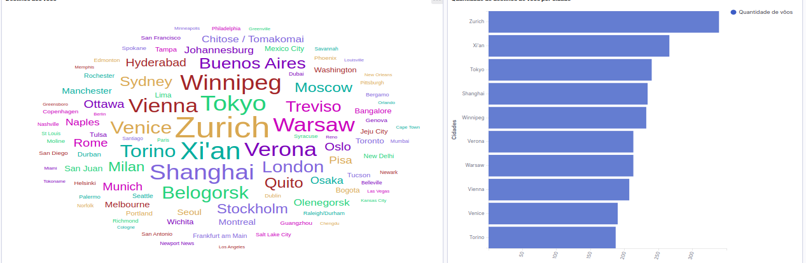
Figura 7 - diferença entre os tamanhos das tags Xi’an e Buenos Aires

Como é possível verificar, as cidades de Xi’an e Buenos Aires possuem um valor máximo de passagem muito próximo, mas a tag Buenos Aires ocupa mais espaço e recebe mais destaque apenas pela quantidade de letras. Essa característica torna as comparações de valores próximos mais difíceis e pode levar a conclusões erradas.

Portanto, a nuvem de palavras não permite uma análise quantitativa mais efetiva dos valores das métricas, pois é possível notar que, nessa nuvem de palavras, o valor real não é exibido, diferente do gráfico de barras, que exibe o valor em um dos eixos. Desse modo, as nuvens de palavras são usadas, principalmente, por questões estéticas e por darem uma noção geral sobre valores associados a uma grande quantidade de categorias, embora pouco efetiva.

Porém, o que fazer quando for necessário representar muitas categorias de forma precisa? Bem, uma alternativa é usar gráficos de barras que mostrem apenas as seleções dos valores, como os 10 primeiros ou os 10 últimos. Desse modo, observe que há muita perda de informação sobre os valores do meio. Como muitas vezes ocorre em visualização de dados, não existe solução perfeita.

Usualmente, quando ganha-se em generalidade, perde-se em precisão, e vice-versa. Portanto, é preciso pensar no objetivo da visualização, na informação que você deseja transmitir ou explorar, e escolher a opção que mais se adequa. Porque, em muitos casos, pode ser vantajoso combinar visualizações em um dashboard, como uma nuvem de palavras para mostrar uma noção geral dos dados, e o gráfico de barras de uma seleção para mostrar detalhes específicos, conforme ilustrado a seguir:

Figura 8 - combinando nuvem de palavras com gráfico de barras em um dashboard

Caro aluno, você chegou ao fim deste conteúdo! Ao decorrer dele, foi possível explorar o funcionamento da nuvem de palavras, assim como criar e customizar esse gráfico, utilizando o Kibana. A partir desses conhecimentos adquiridos, você pôde fazer comparações com o gráfico de barras, além de conhecer estratégias de quando utilizar cada um.

Houve também um momento de discussão sobre a precisão da visualização das nuvem de palavras, entretanto foi possível concluir que elas são esteticamente atrativas, quando há muitos termos, permitindo uma noção bem geral dos termos mais relevantes. Ou seja, uma ferramenta útil de se ter no seu leque de opções de visualização de dados.

Por tudo isso que foi ensinado, é sempre importante lembrar que, em visualização de dados, é necessário combinar estética e funcionalidade para criar gráficos que cumpram bem o seu objetivo, seja para comunicar padrões descobertos para um público determinado, ou ajudar um cientista de dados a compreender e desenvolver um conjunto de dados.

Espero que você tenha aprendido e explorado bastante. Até a próxima!

* Kibana Guide, <https://www.elastic.co/guide/en/kibana/current/index.html> - Acessado pela última vez em Junho de 2020.
* MUNZNER, Tamara. **Visualization Analysis & Design**. CRC Press, 2014.
* WILKE, Claus O. **Fundamentals of Data Visualization**. O’Reilly, 2019.